Computer Ethics - Philosophical Enquiry (CEPE) Proceedings

Volume 2019 *CEPE 2019: Risk & Cybersecurity* Article 7

5-29-2019

Big Data and the Reference Class Problem. What Can We Legitimately Infer about Individuals?

Catherine Greene
London School of Economics and Political Science

Big data and the reference class problem: What can we legitimately infer about individuals?

Catherine Greene
London School of Economics and Political Science

Abstract

Big data increasingly enables prediction of the behaviour and characteristics of individuals. This is ethically concerning on privacy grounds. However, this article discusses other reasons for concern. These predictions usually rely on generalisations about what certain sorts of people tend to do. Generalisations of this sort are often under scrutiny in legal cases, where, for example, lawyers argue that people with prior convictions are more likely to be guilty of the crime they are currently on trial for. This article applies criteria for distinguishing acceptable from unacceptable generalisations in legal cases to a number of big data examples. It argues that these criteria are helpful, and highlight three issues that should be taken into account when deciding whether predictions about individuals are ethical.

Keywords: Big data, ethics, generalisation, reference class, target class.

Big data allows us to characterise, and predict, the behaviour of groups of people. However, increasingly, big data also makes it possible to make predictions about individuals. Recent examples include targeting surveillance at streets, gangs, or individuals who are statistically more likely to commit a crime (Zwitter, 2014), and prediction of individual's pregnancy status by Target, before these individuals had announced their pregnancies. (Crawford and Schultz, 2014). Crawford and Schultz argue that such uses of data can cause "predictive privacy harms" (2014, 93) and argue that harms resulting from predictive privacy are not satisfactorily covered by existing data privacy laws, primarily because, as in the Target example, the company had never collected any data that showed that a particular customer was pregnant, but predicted this information from aggregate data. Their suggested remedy is to institute a procedural policy to mitigate harms.

This article makes a different proposal by challenging the ethical status of predictions about individuals which are based on their membership of a reference class. Drawing on literature discussing the use of reference class data in legal cases (Dahlman 2017) it argues that users of big data have an obligation to analyse reference classes in detail before ethically permissible inferences can be made about individuals. This article discusses four examples, the first is an imaginary one inspired by the film Minority Report, the second example predicts the performance of juvenile criminals in rehabilitations programs, the third predicts future health conditions, and the fourth predicts income levels based on social media posts. This article argues that the ethical status of these prediction can be questioned even before privacy, or even discriminatory issues, are raised.

The argument begins by demonstrating that generalisations tie reference classes to target classes. In other words, that a generalisation of the form 'people who live in a deprived neighbourhood are more likely to commit crimes' underpins the belief that a person belonging to the relevant reference classes in likely to

behave in a certain way. Dahlman proposes eight questions to assess generalisations. Whether the generalisation is true or false, whether the generalisation is robust, whether the generalisation triggers bias, whether the reference class is heterogeneous or homogeneous, whether there is a risk that the generalisation, if accepted, will be overused or overestimated, whether the generalisation is discriminating, and whether the generalisation will put people in the reference class at an unfair advantage (Dahlman, 2017, pg. 98). When these are applied to the four examples they allow us to differentiate the ethical status of different reference class data. The article concludes that researchers seeking to make predictions about individuals using big data have an ethical obligation to assess the quality of the data they are using by paying attention to the truth and robustness of these generalisations.

Predictions about individuals

The collection of large amounts of data increasingly enables companies and governments to make predictions about individuals. In many cases, these predictions are based on large data sets, which make predictions based on what other, similar, people did in similar circumstances. Additionally, much of the information on which such predictions are made is not, obviously, information that people want to keep private.

Perry *et al.* cite research suggesting that short term negative life events can increase criminal activity. These life changes include losing a job, drug or alcohol problems, fighting with a spouse or partner. (2013, pg. 91). The predictions about individuals were based on their similarity to offenders whose history prior to their criminal activity was known. Crawford and Schultz note that such predictions about individuals are enabled through the aggregation of publicly available information.

Big Data's ability to discriminate while manoeuvring around privacy regulations comes from its methodology. Not only can massive amounts of online behaviour be collected and assessed to compute the probabilities of an individual's particular demographic characteristic, but that predictive analysis can also become a form of PII (personally identifiable information) itself. Moreover, this process can predict highly intimate information, even if none of the individual pieces of data could be defined as PII. Although these predictive processes may generate an inaccurate characterization, such processes nevertheless create a model of possible personal information and associate it with an individual. (2014, pg. 101)

There are two possible responses to such predictions. The first is to expand existing privacy regulations; the second is to implement an ethical framework for making predictions about individuals that is not tied to privacy. The advantage of the second approach is that it prevents users of data assuming that 'anything goes' once they have satisfied privacy requirements. This paper will argue for the second approach, while noting that this needs to be underpinned by stringent privacy regulation. The following section presents four examples of predictions about individuals that will be discussed in the remainder of the paper.

1. Predicting future criminals

This example is hypothetical, but gets at the heart of many concerns about making predictions about individuals, partly because of the popular appeal of the film *Minority Report*. In this example, future criminality is predicted on a basis of factors

including employment status, living in deprived neighbourhoods, past convictions, educational level, and being male. Zwitter comments that, “While the future might not be as bad as depicted in the movie, ‘predictive policing’ is already a fact in Los Angeles, where Big Data analytics point to certain streets, gangs or individuals, who are more likely to commit a crime, in order to have them subjected to extra surveillance.” (2014, pg. 4).

2. Predicting juvenile reoffending

Perry *et al.* describe a number of examples, including the prediction of reoffending by juvenile criminals. The Florida Department of Juvenile Justice’s Bureau of Research and Planning manages the Juvenile Justice Information System (JJIS). This system collates the criminal histories of more than 1 million juvenile criminals, and is used to predict the success of juveniles placed in rehabilitation programs. In 2007, the Department piloted a statistical profiling system to predict the juvenile criminal behaviour of 85,000 youths and assign them to risk specific rehabilitation programs in an attempt to reduce recidivism. The youths were assigned to specific programs based on the experience assigning other, similar youths to such programs in the past. These programs included placement with sponsor families, community members, educational opportunities, disciplinary help and vocational training. Perry *et al* note that the programme has not been scientifically tested, but six months after the initiation of the programme, 93% of the youths remained arrest-free (2013, pg. 109-110).

3. Predicting future health conditions

Ding *et al.* (2016) have developed an algorithm for predicting individuals’ future health conditions using data from insured inpatients in China. The aim of these predictions is to intervene early to reduce the incidence of serious disease. The variables collected included demographic information such as age, gender, rural vs urban address etc, and medical information including costs associated with the hospital stay, the diseases suffered from, and length of hospital stay. They used a training set to calibrate the model and then potential diseases a particular patient would get in the future was predicted. They note that, “if patients have several diseases in common, they are considered more *similar* than others. Among *similar* patients, one’s condition can always be taken to evaluate another’s. But if patients share some common diseases such as influenza, they should not be taken as *similar*.” (2016, pg. 759).

4. Predicting income

Matz *et al.* develop a model to predict individuals’ income levels from Facebook profiles. They note that information about peoples’ income is useful for marketing and recruitment. However, it is considered private, and most of us are unwilling to share it. Matz *et al.* show that income is predictable from ‘likes’ and status updates on Facebook. Their analysis suggests that the ‘likes’ of high and low socio-economic status individuals varies markedly- those of high income individuals refer to expensive travel and retail brands, while those of low income individuals refer to luxury in a more generalised way. They also found differences in the use of language in status updates. They do note that approximately 75% of the variance in income remained unexplained by their model, their model was accurate enough to be useful for marketing purposes. The model was, unsurprisingly, better at predicting income levels at the extremes, rather than distinguishing between middle income levels.

Privacy harms

The usual way to approach such predictions is to argue that privacy has been violated. Recent examples include Crawford & Schultz (2014), who argue for a "procedural data due process" (2014, pg. 93). They note that by collating publicly available information, we can predict other data about individuals. The predicted data can be data we would normally consider private. This is the case with the Facebook example above. They note that federal regulations prohibit discrimination in access to credit, but by analysing online activity, such as 'liking' certain status updates loan companies are able to target people likely to have particular credit profiles. They describe this as the creation of "surrogate" data (2014, pg. 99). Companies doing this are able to circumvent existing privacy regulations because, although it is possible to predict highly sensitive information, none of the individual pieces of information on which such predictions are based are PII. Furthermore, in advance of making predictions, it is impossible to know which data will be critical.

Crawford and Schultz propose regulating "the fairness of Big Data's analytical processes with regard to how they use personal data (or metadata derived from or associated with personal data)" (2014, pg. 109). For example, if a hospital patient is judged at high risk of a particular medical condition, and is refused health insurance on that basis, that patient would have a data due process right with regard to that decision. The patient should have access to a due process hearing, where they can present evidence of the harm the decision caused them. They note that, over time, judicial and legal oversight would make such tribunals more efficient. Their approach also requires those who use big data to disclose publicly the types of predictions they attempt, the general sources of data upon which they draw, and make this transparent to the people whose data may be collected. These disclosures should provide access to an audit trail created in the predictive process. Once this disclosure has been made, the people affected should have access to a hearing, where the fairness of the predictive process can be challenged.

The following section seeks to extend this approach, by providing criteria which allow users of big data to assess the ethical status of predictions they are seeking to make before they cause actual harms. The first step is to outline the structure of predictions about individuals.

Predictions about individuals are made on the basis of membership of reference classes. Big data algorithms find correlations in large data sets. For example, a data set of purchasing history at a supermarket might find that people who buy a particular brand of shampoo go on to buy a certain brand of hairbrush, with a certain probability. Using this information, we can predict that a particular customer, who has bought that shampoo has a certain probability of going on to buy the hairbrush. The supermarket may then decide to market this brush to them, for example by showing them adverts when they shop online. For the purposes of the following discussion, the following definitions will be helpful:

> Prediction: That person P, who has bought Shampoo X, is likely to buy Hairbrush Y.
> Generalisation: People who buy Shampoo X, buy Hairbrush Y with probability *a*.
> Population data set: All customers of the supermarket.

Reference class: Customers buying Shampoo X.
Target class: Customers buying Hairbrush Y.

Person P is in the population data set. The prediction about their future purchase (membership of the target class) is based on their membership of the reference class, and the generalisation about the behaviour of other people in that reference class. This is the same form of argument in the cases above. In the criminality example the prediction of future criminals is based on their membership of a reference class of people who committed crimes, along with a generalisation that people with these characteristics commit crimes. In the juvenile intervention example, the prediction about which juvenile intervention will work for a particular young offender is based on their membership of a reference class for whom this intervention worked (or worked better than other alternatives) in the past. In the second example, the prediction about future health conditions in based on that individual's membership of a reference class of other similar patients, and the health conditions that they went on to suffer. In the third example, the prediction of an individual's income level is based on their membership of a reference class of other individuals liking similar topics, and making similar comments, on Facebook.

This form of reasoning is analysed in the literature on legal evidence. The following section reviews this, before arguing that the ethical restrictions on using such reasoning are applicable to the big data cases.

Legal principles and generalisations

Dahlman argues that arguments about legal evidence often rely on generalisations, only some of which are acceptable. An argument about legal evidence points to this evidence and suggests that it increases the probability of a certain hypothesis, usually the guilt or innocence of the defendant. For example, the defendant is recorded on CCTV close to the scene of the crime. This increases the probability that the defendant is guilty. This argument depends on a generalisation that links the evidence with the hypothesis. In this case, the generalisation that people close to the crime scene are more likely to have committed the crime. Dahlman refers to a reference class and a target class, in the same way as the example above.

> As we shall see, a generalization connects two classes to each other. I will refer to these classes as the ''reference class'' and the ''target class''. When an argument on legal evidence points to a certain piece of evidence, it classifies the case at hand as belonging to the reference class of cases where this kind of evidence is present, and when the argument claims that the evidence increases the probability of a certain hypothesis, it claims that membership in the reference class increases the probability that the case belongs to the target class of cases where the hypothesis is true. (2017, pg. 84)

Dahlman then considers what it is about these generalisations that make them unacceptable, and what it is that distinguishes them from acceptable generalisations. He provides a checklist of "critical questions" (pg. 98) that help to make this distinction.

1. Is the generalisation empirically true, or false, as a generalisation?
   Is membership in the target class more common in the reference class than among cases in general?

2. Is the generalisation sufficiently robust?
   Is the reference class homogenous or heterogeneous?
3. Does the generalisation trigger bias?
   Is there a risk that the generalisation, if accepted, will be overused, or overestimated?
4. Is the generalisation discriminating?
   Does the generalisation put people in the reference class at an unfair disadvantage? (2017, pg. 98)

Is the generalisation true?

Consider the following case: A witness in a burglary case testifies that they saw a man loading boxes into a van. However, it was night-time so he couldn't identify the man, but thought the van was blue. In his summing up, the defence attorney says that it is common knowledge that colours are easily mistaken in the dark. Blue might easily be mistaken for green, and vice versa. The witness might therefore be mistaken. Dhalman notes that this sort of generalisation is usually acceptable.

The generalisation in question is: "the probability that an observation is mistaken, given the information that it was made in the dark, is higher than the probability that it was mistaken, given that we are ignorant about the light conditions when the observation was made. For this to be true *membership in the target class must be more likely in the reference class than in cases in general*." (2017, pg. 89. Italics in original). In judging the truth of this generalisation it is not sufficient that the observation was made in the dark, it must be true that mistakes are more common when made in the dark, than when we do not know the lighting conditions. This is because some observations will be mistaken, regardless of light conditions. For this generalisation to be true, errors need to be more common in the dark. To generalise this point, it is not sufficient that membership of the target class is common in the reference class, it must be more common in the reference class than in the population in general. This generalisation is therefore relatively unproblematic because mistakes about colour are more common in the dark.

Is the generalisation robust?

Consider the following case: A man is standing trial for the murder of his neighbour with a shotgun. The man claims that his wife shot the neighbour. The prosecutor argues that, because only 8% of homicide offenders are women, it is probable that the man killed the neighbour, rather than the woman.

This argument relies on the generalisation that it is more probable that defendant is guilty if they are a man. Dahlman notes that this is correct, statistically speaking. However, this generalisation is concerning, despite being true. Essentially, this particular man is being judged on the basis of his membership of a group- being a man. Suppose that the defendant has a track record as a peaceful and law-abiding citizen. In this case it is more unacceptable to judge him on the basis of being a man, than on the basis of being a man, with a peaceful and law-abiding history. This is because, Dahlman argues, 'being a man; is a more heterogeneous reference class than 'being a man with a peaceful and law-abiding track record'. There is more variability in the group, 'men', than in the group 'men with a peaceful and law-abiding

history'. A generalisation is more robust, and more acceptable, when the reference class is more specific. In this case, it is unacceptable to use this generalisation because 'being a man' is too heterogeneous a reference class to be relevant in this case.

Does the generalisation trigger bias?

In some situations we might be reluctant to accept a generalisation despite its truth and robustness because it has the potential to bias. Consider the following case: A man of Somali origin is standing trial for shoplifting. The prosecutor uses crime statistics showing that convictions for shoplifting are more common among people of Somali origin than for the population in general to argue that, although these statistics do not mean this particular man committed this offence, they do increase the probability that he did.

The worry with this generalisation is that it will be overestimated, or over-used. Dahlman notes that data showing that shoplifting convictions are higher for men of Somali origin than for men of other ethnic origins does not mean that men of Somali origin are overrepresented among guilty defendants. Their higher representation among those convicted for shoplifting could be due to bias against them in the legal system. Using this generalisation could perpetuate and exacerbate any existing bias, and might even legitimise racism in the population. Furthermore, it is difficult to judge the extent to which such statistics, even if true and not attributable to bias, increase the probability that this particular defendant is guilty. There is a risk that this generalisation will trigger bias against the man on trial.

Is the generalisation discriminating?

This is an additional worry with the Somali example, because it discriminates against people from a particular ethnic origin. Accepting this generalisation has potentially large negative effect for other people from a Somali background. Dahlman says that they may be "systematic disadvantage" (2017, pg. 97). Dahlman notes that many generalisations can be discriminating. Consider the case where a mother provides her son an alibi. The prosecution might argue that mothers are unreliable providers of alibis because they are likely to lie to protect their children, particularly if their children are on trial for serious crimes. The generalisation that mothers are likely to lie to protect their children is discriminating against mothers, if it is accepted. However, Dahlman argues that it is less discriminating because the negative repercussions for mothers are less than the negative repercussions for people of Somali origin. Mothers providing truthful alibis for their children (and those children) may well rankle at this suggestion. However, Dahlman is correct to note that when generalisations discriminate, the seriousness of this discrimination must be weighed against the evidential value they provide. The generalisation about Somali origin is discriminating, and therefore unacceptable.

Dahlman does not delineate the difference between bias and discrimination in quite these terms, but in the following discussion bias will refer to negative effects on individuals, while discrimination refers to wider effects on other people classified as belonging to the same group. We can therefore distinguish between bias and overuse against an individual, and wider discrimination against a group of people.

Summary

Dahlman provides criteria that help to decide whether generalisations are acceptable, or unacceptable. These work relatively well for his legal examples. The following section applies them to the big data examples introduced at the beginning of this paper. It shows that these criteria are transferable to the big data cases.

How do these apply to the big data examples?

Example 1: Predicting criminals.

Is the generalisation true?
This example is hypothetical, so the truth of the generalisation is difficult to judge. However, there is a clear problem with this case. This is because what we really want to know is whether membership of the target class is higher in the reference class than in the population in general. However, not all crimes are discovered, reported, or recorded. Particularly for less serious crimes, we do not know the extent of criminality in the population at large. This makes generalisations linking reference classes with target classes problematic. Any statistical relationship between the reference and target classes is likely to be untrue because we will not be sure that membership of the target class is higher in the reference class than in the population in general. An additional worry is the incidence of false convictions. Any statistical relationships based on characteristics shared by convicted criminals will include the characteristics of wrongly convicted individuals. By its very nature, the incidence of false convictions is difficult to judge.

Furthermore, in this case, we also need to know how many convicted criminals do not have the characteristics defining the reference class. What percentage of criminals share characteristics with the reference class? Barocas *et al.* (2016) discuss this problem in the context of employment practices. Obtaining data with sufficient granularity can be expensive, and sometimes does not exist. They note that "Obtaining information that is sufficiently rich to permit precise distinctions can be expensive. Even marginal improvements in accuracy may come at significant practical costs and may justify a less granular and encompassing analysis." (2016, pg. 689). There may simply be insufficient data to be able to understand the extent to which the target and reference classes match.

Finally, a system that finds correlations in data will often not find causal relationships. However, if we want to use correlations to make predictions about people causation is relevant, and makes a difference to the acceptability of these generalisations. For example, if the relationship between age and criminal convictions in causal, rather than just a correlation, it may be more ethically permissible to use it.

Is the generalisation robust?
The reference class is robust to the extent that it is less heterogeneous. However, the issue is not just about the degree of heterogeneity *per se*, but what this heterogeneity is about. For example, it makes little difference to the robustness of

the generalisation whether the reference class is heterogeneous because it includes characteristics such as, a low income, poor educational attainment, liking black trainers, eating at McDonald's, and watching American Idol. To be judged more heterogeneous, the characteristics present in the target and reference class must be 'relevant'. Judging relevancy is difficult, but this can also be seen as an issue of distinguishing causally relevant factors from correlational ones.

Does the generalisation trigger bias?
This generalisation has potential to trigger bias, both because people in the reference class may be judged more likely to commit crimes than they really are. Additionally, Anderson (2012), in his discussion of character evidence in trials, highlights a risk that individuals are convicted on the basis of the sort of person they are, rather than because of what they did. This worry is relevant here because juries, judges, and the police may become biased against people 'of a certain kind' because of the perceived connection between certain characteristics and criminality.

Is the generalisation discriminating?
Anyone in the reference class is open to discrimination, from other members of the public, and from the police. People in this group might be arrested speculatively, and might suffer discrimination in job applications. Barocas and Selbst (2016) raise a wider worry about discrimination that can arise in cases like this. They note that not all data is collected 'equally'. Certain citizens, or neighbourhoods are often overlooked (2016, pg. 684). They write that this can arise due to lack of access to technology, through which data is often collected, or because people do not participate in the formal economy. They illustrate this with an example about pothole monitoring. In Boston, potholes could be automatically reported through a smartphone app which detected when a car drove over a pothole. Given the differential rates of smartphone ownership, this reporting mechanism is biased towards those in affluent neighbourhoods. This could also be a problem in the predictive policing example. There may be differential rates of crime reporting, or recording, or even in conviction rates for crimes in different areas. Even if the generalisations use were true, and robust, the underlying data may not accurately reflect the real distribution of crime across different neighbourhoods. In an extreme case, the data might be used to predict people likely to commit crimes in affluent neighbourhoods, rather than all neighbourhoods.

Summary
Predicting future criminals using big data is unethical on each of the criteria above. The generalisation isn't true because we are unlikely to have complete population data to use for analysis. The population data is likely to be more reliable for serious crimes, such as murder, but not even all murders are discovered, or reported. Additionally, not all those convicted are guilty. These issues make it difficult to specify the reference class correctly. This generalisation may not be robust, and raises significant concerns about bias and discrimination. It is therefore unacceptable. It is important to think about the data that we want—the incidence of criminality in the population at large, not just the data that we can get. An awareness of the limitations of the data should encourage an appreciation of the limitations, and possible unreliability, of generalisations.

Example 2: Juvenile reoffending

Is the generalisation true?
The research underlying this case has not been scientifically validated, however, the initial evidence is that it is true. The worry about reference classes is not so important in this case. The population is juvenile offenders. The reference class is juvenile offenders who have had access to a number of different interventions, and the target class is juvenile offenders who share various characteristics with offenders in the reference class. What we need to know, in this case, is not only that juvenile offenders with a certain group of characteristics have responded well to a particular intervention, but that they have done better than juvenile offenders in general. The research does not show this, but does show a reduction in reoffending levels. In advance of knowing this, is it unclear whether this generalisation is true, in Dahlman's sense.

Is the generalisation robust?
The reference classes may be more, or less robust. In principle, it is possible to refine them so that they are less heterogeneous. However, decisions about how much data to collect may affect the degree of homogeneity of reference classes, and should be considered an ethical decision. Zook *et al.* (2017) propose 10 guidelines for ethical 'big data' research. The first four focus on privacy and the use of seemingly innocuous data that can be used to discriminate against people. The fifth encourages researchers to "Consider the strengths and limitations of your data" (2017, pg. 4). They write that the context within which data is gathered is important for the interpretation of that data. Rather than suggesting clear conclusions, data is often unclear, and subject to interpretations. This is particularly important, in the current example; young offenders may be classified in ways that miss unique factors that may affect their rehabilitation. For example, a number of juvenile criminals could belong to a gang, but this may be for very different reasons. For some youths the gang may be a surrogate family and provide social support which is lacking elsewhere in their lives. For others the gang may facilitate criminal behaviour they were engaging in before they joined the gang. Classifying them all as 'gang members' does highlight a feature they have in common, but may disguise a great deal of heterogeneity that could be relevant for their rehabilitation. The robustness of the generalisation therefore needs to be carefully considered.

Does the generalisation trigger bias?
There is some potential to trigger bias if some juvenile offenders are categorised as particularly troublesome, or problematic. Additionally, some bias might be triggered if some offenders are judged 'not capable' of participating in educational interventions. However, these concerns deal with the implementation of the interventions, and the way in which the reoffending interventions are managed, rather than with the generalisation itself.

Is the generalisation discriminating?
The population of juvenile offenders is, presumably, already at a disadvantage because they have committed crimes. If the programs limit reoffending then this has the potential to reduce discrimination. If the wider population learns that young offenders can move on to live fulfilling lives after such interventions then this

generalisation has the potential to reduce discrimination. This might particularly be the case in employment situations.

Summary
The data is being used to allocate juvenile offenders to a variety of positive interventions. The case would be very different were it used to decide which juveniles should have access to these interventions at all. As such, this use of big data to make predictions about individuals is potentially acceptable. However, it is important to ensure that juvenile offenders do not experience any bias in the allocation to one scheme rather than another. It is also important to consider the extent to which the generalisation is robust.

Example 3: Health conditions

Is the generalisation true?
The relationships discovered in the cited study appear relatively significant. However, the issues discussed here also relate to a more general 'prediction of disease' framework. The truth of such generalisations depends on knowing the incidence of particular diseases in the wider population, which depends in turn on the accuracy of detection and diagnosis at a population level. Char *et al.* (2018) provide a number of examples of 'missing' data in health care datasets. This includes cases where few studies have been done in certain populations; they write, "Attempts to use data from the Framingham Heart Study to predict the risk of cardiovascular events in non-white patients have led to biased results" (2018, pg.1). A different problem resulting from missing data is found in big data analysis of mental health. Chancellor *et al.* (2019) describe how social media posts have been used to predict which users are suffering from depression, and other mental health issues. The dataset used was from people who had already disclosed a mental health condition on social media (a post saying 'I've been diagnosed with depression' for example). Matching the characteristics of these people to potential sufferers from depression misses the fact that the sample of people suffering from depression who are comfortable disclosing this on social media is a subset of those with depression. Predicting depression based on the characteristics of this subset is likely to miss a large number of relevant people. (Chancellor *et al.* discuss other worries with using this data that question whether it is ethically permissible for other reasons). The truth of these sorts of generalisations needs to be carefully considered to understand whether the data available is characteristic of the population, the reference class, and the target class researchers would like to know about.

Is the generalisation sufficiently robust?
It is likely that the range of factors that might be relevant to the incidence of disease is more restricted than in the criminality example. These sorts of data sets have the potential to be more robust, but this will vary for different health conditions.

Does the generalisation trigger bias?
Generalisations of this sort could easily trigger bias, if treatment is affected, or if insurance premiums are raised for specific individuals. For example, treatment might be rationed away from patients expected to suffer from additional health conditions in the future, and insurers might raise premiums for patients judged at high risk of

future health conditions. Furthermore, suppose that a patient is instructed to change their lifestyle, such as taking more exercise or making dietary changes to reduce their chances of suffering from a medical condition. If they then still suffer from the medical condition a few years later, there is a risk that they will suffer negative bias from the medical community, or from the wider community, as they are seen as bringing about their own decline.

Is the generalisation discriminating?
Some medical conditions may be causally related to living in poverty, lifestyle, or genetic factors. There is the potential for people from such wider groups to be discriminated against by the medical profession, the wider community, and employers. There is also a risk that insurance companies charge these communities higher premia, or become reluctant to ensure them, even if they are disease free at the time a policy is taken out.

Summary
The purpose of these predictions about individuals is to reduce the incidence of disease, and to help people by intervening before they develop serious health conditions. This is an ethically praiseworthy desire. The positive aspects of these predictions need to be isolated from the possible negative consequences. In cases like this, implementation and management is critical. Perhaps, patients should be required to opt-in to predictive programmes, and data privacy would be especially important is such cases. Additionally, the predictive policing and medical examples suggest that decisions about whether to think about causation is an ethical one. For example, if lifestyle factors are judged predictive of future health conditions, and the treatment of patients is affected by these lifestyle choices it is relevant to know whether these are causal, or merely correlational. Zwitter *et al.* (2014) note that the more data we have, the more likely it is that random commonalities will be found. They write that “in fact, no connectedness at all would be the outlier.” (2014, pg. 5). On a purely pragmatic level, correlational relationships may be less likely to persist than causal ones. Johnson (2014) provides a relatively benign example. A US college discovered a relationship between first day login and online course success. The college assumed this relationship was causal and encouraged students to log in on the first day, expecting that this would increase their success on the course. It had no effect. Although the intervention did no harm, resources could have been used more productively. Distinguishing causation from correlation is important when attempting interventions that may affect people’s health. Additionally, the availability of data may affect the truth of these sorts of generalisations.

Example 4: Income levels

Is the generalisation true?
This example can be generalised to cover many different sorts of predictions based on social media use. The population data set in these sorts of cases is all users of certain social media, in this example Facebook. The reference class is people with various levels of income matched with various characteristics on social media. The target class is people who share those social media characteristics, and whose income is then predicted. These generalisations will be true if membership of a target class specifying a particular level of income is higher in the reference class than in

the population in general. As suggested by the research into income levels, these generalisations are more likely to be true at extremes- very high and low income levels, than for a wide range of average income levels.

Is the generalisation sufficiently robust?
The generalisation could be very robust if users reveal a lot of information on social media, which would allow for the construction of accurately specified reference classes. However, there are reasons to worry about the robustness of these sorts of datasets. For example, Zook *et al.* (2017) ask, “Are your findings as clear-cut if your interpretation of a social media posting switches from a recording of fact to the performance of a social identity?...it is fundamental that researchers be sensitive to the potential multiple meanings of data.” (2017, pg. 11). They go on to note that a Facebook post could be interpreted as approval or disapproval of something, a simple observation, or an attempt to improve social status. Attending to these nuances affects the hypotheses considered, as well as the conclusions drawn from analysis. Data sets may be more heterogeneous than they appear. For example, a number of people may ‘like’ the same status updates, but for different reasons entirely. Seeing this group as similar with regards to this status update may therefore be misleading. The social media example is therefore more problematic than it at first appears.

Does the generalisation trigger bias?
These generalisations could trigger bias if the information is used to deny credit, or raise interest rates on debt, or to market questionable products to individuals. For example, very low income social media users could be shown adverts for poor quality rental properties, rather than higher quality ones they are judged (not necessarily accurately) unable to afford. High income users could be targeted by companies charging more for services that are available at a lower cost elsewhere, because they are judged able to afford it. There is a clear potential for bias against these groups.

Is the generalisation discriminating?
This generalisation could be discriminating. The study cited above noted that both the subject of ‘likes’ and the language used are indicative of income level. Were these generalisations to be come widely known, the use of particular language, or personal preferences might become correlated with characteristics such as ‘being poor’. This could lead to discrimination against people because of the things they say, or write, in the real world.

Summary
Generalisations of this sort may well be true, however there may be significant issues with robustness because people may be doing different things with a social media post. A ‘like’ could be indicative of a variety of different things; approval, an expression of solidarity with friends, or an expression of an online identity. Generalisations may also lead to bias and discrimination.

Conclusion

Dahlman's criteria allow us to raise a number of concerns with these big data cases that go beyond privacy. Bias and discrimination are relatively easy to appreciate, although judgements about them are sometimes arguable. This is illustrated by Dahlman's reduced concern about discriminatory generalisations about alibi providing mothers than about people of Somali origin. Dahlman's criteria are useful for highlighting additional ethical concerns that may preclude the use of some generalisations even before we consider bias and discrimination. These are concerns about the data researchers use in their analysis. If acceptable generalisations are true and robust, then these three ethical issues are involved in the collection and use of data too:

Ideal vs Actual Data: When making predictions about individuals on the basis of big data analysis, researchers should critically assess the generalisations they are relying on. They should ask what data they need to assess the truth of these generalisations. Relying on a generalisation when there is insufficient data to know whether it is true is unethical. This criteria is particularly important in the predicting criminality and healthcare examples, where missing or unavailable data should throw doubt on the truth of the generalisations.

Causation: Even if it isn't possible to know whether relationships discovered through big data analysis are causal or correlational, this lack of certainly should be incorporated into the confidence attached to these relationships, particularly where potential for bias or discrimination exist. This criterion is important in the predicting criminality, healthcare and juvenile rehabilitation examples, where correlations may be discovered that are not helpful for initiating successful interventions.

Heterogeneity: When making predictions about individuals on the basis of big data analysis, researchers should consider the nuances in the data they use to assess the extent to which people classified as being 'the same' really are the same. This is relevant in all four examples. Data may be classified in ways that are convenient, or easy, but this may miss nuances and differences within classifications that reduce the robustness of a generalisation.

This paper has applied Dahman's criteria for assessing the acceptability of generalisations in legal arguments to predictions about individuals on the basis of big data analysis. It began by showing that the form the arguments take in legal and big data cases is the same, before applying Dahlman's criteria to the big data cases. In addition to concerns about bias and discrimination, this analysis has enabled the formulation of three data related criteria that should be addressed when making predictions about individuals.

References

Anderson, B.J. (2012) *Recognizing Character: A new perspective on character evidence* The Yale Law Review 121.7. 1912-1968

Barocas, S.; Selbst, A. D. (2016) *Big data's disparate impact* California Law Review 671-732

Chancellor, S. *et al.* (2019) *A taxonomy of ethical tensions in inferring mental health states from social media* Conference on Fairness, Accountability, and transparency (FAT 2019), ACM, New York

Crawford, K.; Schultz, J. (2014) *Big Data and Due Process: Towards a framework to redress predictive privacy harms* Boston College Law Review 55.93 93-128

Dahlman, C. (2015) *The felony fallacy* Law, Probability and Risk 14. 229-241

Dahlman, C. (2017) *Unacceptable generalisations in arguments on legal* evidence Argumentation 31 83-99

Ding, R.; Jiang, F.; Xie, J.; Yu, Y. (2017) *Algorithmic prediction of individual diseases* International Journal of Production Research 55.3. 750-768

Johnson, J.A. (2014) The ethics of big data in higher education International Review of Information Ethics 7. 3-10

Matz, S.c.; Menges, J. I.; Stillwell, D. J.; Schwartz, H. A. (2019) *Predicting individual-level income from Facebook profiles* PLoS ONE 14.3

Perry, W. L. *et al.* (2013) “Using predictions to support investigations of potential offenders” in Perry, et al. *Predictive Policing*, RAND Corporation pg 81-113

Richards, N.M.; King, J.H. (2014) *Big data ethics* Wake Forest Law Review 49 393-432

Zook, M.; Barocas, S. et al. (2017) *Ten simple rules for responsible big data research* PLoS Computational Biology 13.3. 1-10

Zwitter, A. (2014) *Big data ethics* Big data & Society 1-6